EXPLOITING RULES FOR RESOLVING AMBIGUITY IN MARATHI LANGUAGE TEXT

Gauri Dhopavkar [1,4], Manali Kshirsagar[2], Latesh Malik[3]

[1] Research Scholar, Department of CSE, GHRCE, Maharashtra, India
[2] Professor, Department of CT, YCCE, Maharashtra, India
[3] Professor, Department of CSE, GHRCE, Maharashtra, India
[4] Faculty, Department of CT, YCCE, Maharashtra, India

Abstract

Natural language ambiguity is a situation involving some words having multiple meanings/senses. This paper discusses natural language ambiguity and its types. Further we propose a knowledge based solution to resolve various types of ambiguity occurring in Marathi language text. The task of resolving semantic and lexical ambiguity occurring in words to obtain the actual sense is referred as Word Sense Disambiguation (WSD). Marathi language is the official and commonly spoken language of Maharashtra state in India. Plenty of words in Marathi are spelled same as well as uttered same but are semantically (meaning-wise/ sense-wise) different. During the automatic translation, these words lead to ambiguity. Our method successfully removes the ambiguity by identifying the correct sense of the given text from the predefined possible senses available in Marathi Wordnet using word and sentence rules. The method is applicable only for word level ambiguity. Structural ambiguity is not handled by this system. This system may be successfully used as a subsystem in other Natural Language Processing (NLP) applications.

Key Words: *Word Sense Disambiguation, Natural Language Processing, Marathi, Marathi Wordnet, ambiguity, knowledge based*

---***---

1. INTRODUCTION

All living things/organisms need to communicate with each other for their survivor. They do it using different methods. Many communicate through various event specific sounds, gestures etc. Homo sapiens are intelligent among all the species and they share their thoughts through Natural Language. Through natural language, all ideas, thinking, views are communicated accurately. Still, due to some features of language, meanings of words may shift leading to misunderstanding and miscommunications. Even if literary parameters are followed strictly, by virtue of nature, each natural language suffers from Ambiguity.

As per the Oxford dictionary[1], the term “Ambiguity” refers to the state of having or expressing more than one possible meaning or something open to more than one possible meaning. It refers to the state in which any linguistic entity, any symbol, a word or a sentence (statement), any text can be understood in more than one way. For humans, they being intelligent, may overcome misunderstanding and miscommunication caused by language ambiguities, by using various ways of analysis naturally. But getting the jobs done with the help of machine is a complex task as it lacks the knowledge and the common sense reasoning. Ambiguity poses problems in majority of the NLP tasks like Machine Translation, Text Summarization and Named Entity Recognition etc. In order to deal with the problem of ambiguity, it becomes necessary to understand the reasons behind its occurrence, its types and the various levels at which it may occur.

The paper is organized in following sections. Section 1 introduces the concept of ambiguity and its importance in natural languages processing. In section 2, we discuss word sense disambiguation, Section 3 focuses on literature available in field of disambiguation. In section 4, we present our approach to tackle the problem of ambiguity. Section 5 details the conclusion of the work presented by us.

1.1 Types of Ambiguity in NLP tasks

Ambiguity represents the state where it is confusing, unsure to fix a precise meaning. It also becomes much cumbersome to provide an explanation, since it involves different meanings. Unclearness represented by ambiguity is because of having more than one meaning. The basic ambiguity may be a word with multiple senses. For example, consider the word “fly” means “A type of insect or two winged creature” or “to move through air”.

The outcome of ambiguity is the confusion in the mind of a reader in case of written text. Ambiguity also creates unnecessary confusion in hearer’s mind in case of speech. Because of this confusion or uncertainty, no effective communication is possible.

Ambiguities can be classified in different ways depending upon the principles used for classification and the reason behind the occurrence of ambiguity. Ambiguity is classified as:

Intentional Ambiguity (related with any valid literary text.)

Unintentional Ambiguity (related with real world language use).

Many times ambiguity may be considered as local or global depending on the scope of it.

Marathi language philosophy indicates further division of ambiguities as mentioned below: [2]

The ambiguity might occur due to same name given to the whole and its part ambiguity- (e.g. "makaa/corn" etc.)

Act and object ambiguities : The action/act and object have same name. example badagaa (object used for beating/action taken against someone) , taancha aaNaNe (heel or bring in problem).

The raw material and the finished product (e.g. "khasa"/waaLaa/stem of one plant with sweet fragrance used in water).

The name of community or the cast and its member is same. For example, maaLii (gardener/name of caste), panDita(one who performs sacred/holy work of religion).

1.2 Levels of Ambiguity

In literature, ambiguity may be decomposed into different levels as below[3] [4]:

- Phonological Ambiguity- Marathi example, (*“**थुंकू नये**.”(Do not spit) or “ **थुंकून ये**.”(spit and come))*, Utterance sounds same but exactly opposite meaning.
- Lexical Ambiguity (Polysemy or Word Sense ambiguity)- In Marathi language, consider example word showing lexical ambiguity as *“warga”(वर्ग)* (representing result of multiplication of a number by itself, i. e. square or other meaning is class room or class/standard).
- Structural Ambiguity-Consider an example in Marathi, Example: *"taruNa muley aaNi mulii naataka baghayalaa aale."(Young boys and girls came to watch drama).*
- Some other levels of ambiguities are also found in natural language text as mentioned below-
- Transformational Ambiguity[5]
- Scope Ambiguity
- Pragmatic Ambiguity
- Pun

2. WORD SENSE DISAMBIGUATION(WSD)

In order to understand the need of WSD, consider the following example-

नदीचे पात्र मोठे आहे.

(River *basin* is large)

तो या पदासाठी पात्र नाही.

(He is not *eligible* for this post)

In the above example पात्र word is ambiguous. In first sentence it’s sense is interpreted as *River Basin* whereas in second sentence it is representing sense as Eligibility for some post. If the correct sense is not interpreted by the machine, it will create chaotic situation.

The activity of identifying the accurate sense of a word as well as sentence is considered as Word Sense Disambiguation process. If the problem of Word Sense disambiguation is not handled carefully, it may lead to disastrous results in the application.

3. LITERATURE REVIEW

In this section we present literature survey carried out for various efforts done to address WSD problem. From the literature survey it is found that very less work is carried out for Indian languages. In India, around 22 official languages are spoken. But efforts for automated NLP tasks has been initiated for very few Indian languages.

The WSD tasks are generally categorized as -Knowledge based and Corpus Based. We limit our discussion only to understand the work carried out for Knowledge based work:

3.1 Knowledge Based Disambiguation

The disambiguation task relying on knowledge bases uses external lexical resources like dictionaries, thesauri or lexical knowledge base (also considered as Computational lexicon). These methods do not need large amounts of training material as required in Supervised methods.

Machine-Readable Dictionaries (MRDs) are a ready-made source of information about word senses and knowledge about the world, which is required for WSD task. Amsler made its first use and later on it became very popular. In this method, the most plausible sense out of number of available senses is chosen which maximizes the relatedness among all the chosen senses.

Thesauri provide information about relationships among words, most notably synonymy.

Enumerative and generative are the two popularly used classes for constructing Computational lexicons. In *enumerative* approach, the senses are explicitly provided, while in *generative* approach, semantic information associated with given words is underspecified. The generation rules are used to derive precise sense information.

Generative Lexicons work on lexicons. Related senses (i.e., systematic polysemy), are not enumerated but rather are generated from rules which capture regularities in sense creation, like for metonymy, meronymy, etc.

The knowledge based methods were used in the work carried out by algorithms like Lesk[6], Walker[7], Agirre and Rigau’s algorithm using conceptual density[8] etc.

The different approaches discussed in various papers [14] use-

- The selectional preferences and the arguments in which exhaustive enumeration of various properties such as arguments, selectional preference of the argument, description of the properties of the word is required. Based on this the selection preference criteria can be decided.
- Overlap approaches which make use of MRD such as WordNet, thesaurus etc. It first identifies the features of overlap in between the different senses of the word that is ambiguous and also the features of the words that are in the context.

Various drawbacks of knowledge based approach mentioned in the paper [11] are:

- Drawback of selectional restrictions is that it needs a very large amount of database.
- Drawback of overlap based approach includes that the definition in MRD are limited and it rarely takes

distributional constraints of differential word sense in account.

3.1.1 Supervised Disambiguation

This approach is based on a labeled training set and the learning system has a training set of feature encoded inputs and their appropriate sense label (category).
Paper [12] describes various approaches used for word sense disambiguation. The approaches described are supervised approaches.
TRUMP[17] (TRansportable Understanding Mechanism Package) system functions on a "core" of knowledge about language and uses a set of techniques for applying the core knowledge within a stipulated domain, the information about words, word meanings, and linguistic relations.
Yarowsky's[13] algorithm uses minimal or no training data often called as co training data. It uses small set of labeled data with a large amount of unlabeled text to learn a classifier with fine accuracy.
Ng and Lee's[15] approach is a very well known approach that uses a nearest neighbor approach called LEXAS. It uses various knowledge sources like parts of speech of words, morphological structure etc.
[16] Lin developed a supervised approach to disambiguate word sense without the use of a classifier. One of the disadvantage of his approach is that it lacks in modeling specific for every word of the training corpus which effects the accuracy of the algorithm. It highly relies on the syntactic dependencies which helps in capturing the context word.

4. PROPOSED METHODOLOGY

In this section we discuss the approach of WSD used in our work.
For carrying out WSD in any language, initially the list of senses of a word is assumed to be fixed with reference to some dictionary of that language. This demands for the availability of a sense inventory for the words in a language. Till early 1990s, most dictionaries for English and other languages were available only in paper form. Hardly a few were available in electronic form and that too only aavailable for a limited group of researchers. This was the major hurdle in the NLP progress. At this point WordNet entered the scene and was like a boon. It is a public-domain electronic dictionary and thesaurus for different languages freely available for research purposes.
(English WordNet is Created by George Miller and his team at Princeton University, (WordNet (Miller 1990, Miller and Fellbaum 1991, Fellbaum 1998). Based on the idea of English Wordnet, other language Wordnets are also designed. Dr. Pushpak Bhattacharya of IIT Bombay alongwith his team has developed WordNets for Indian languages like Hindi, Marathi etc. Marathi Wordnet is also a Wordnet based of semantic relatedness of words. It is a large electronic database organized as a semantic network. It is constructed on paradigmatic relations involving synonymy, hyponymy, antonymy, and entailment etc. It has become the most widely used lexical database today for NLP research in these languages, WordNet lists the different senses (called synonym sets, or synsets) for each open-class word (nouns, verbs, adjectives, and adverbs).

In our work we have used Marathi Wordnet 1.3[10]. Marathi Wordnet has following files-

Index_txt: Provides information about all words present in Wordnet

Data_txt: Provides details of every word in index file.
Onto_txt : Provides ontology details of the words in data file.
(data is obtained officially from IIT Mumbai) [9, 10]
Structure of Data_txt:
Example:
00054558 03 02 चालणे:धकणे 0001 0400 00000143 | हेतुपूर्ततेला उपयोगी पडणारे असणे:"चूल पेटवायला ओली लाकडे कशी चालतील?"

synset id=00054558
pos=03
number of words present in synset=02
synsets= चालणे, धकणे
number of relations lexical as well as semantic=0001
four digit code relation id=0400
synset_id for which that relation exists=00000143
gloss= हेतुपूर्ततेला उपयोगी पडणारे असणे
example sentence=:"चूल पेटवायला ओली लाकडे कशी चालतील?"

pos: 1(noun), 2(adj), 3(verb), 4(adv)
Antonymy and Gradation relations are represented with the help of two four digit code (first four digit represents relation type and second four digit represents the order of words from two synsets for *which* relation holds)

Consider the following example in index.txt file:
अंक 01 01 0400 05 00002878 00002910 00004049 00010697 00001958
In the above example:
word= अंक
pos= 01 {*pos: 1(noun), 2(adj), 3(verb), 4(adv)}
number of relations exists for word in all its senses=01
number of senses=05
Structure of Onto_txt :
00000043 0001 00000035 | मारक घटना *(Fatal Event) {FTL उदाहरणे :- धरणीकंप इत्यादी}

onto id=00000043
0001 indicates that parent exists
parent onto id=00000035
onto description=(Fatal Event)

In order to parse these files we have designed different parsers like onto_txt parser, index_txt parser and data_txt parser Thus we get complete lexical as well as meaning wise(sense) detail of the text under consideration.

4.1 Algorithm Steps

Our proposed Algorithm performs disambiguation through the following steps:

1. Text or Discourse is Input (ambiguity might be present or might not present)
2. Parse the files like index_txt, onto_txt, data_txt available in Marathi WordNet 1.3 using specially designed parsers.
3. Identify the presence of ambiguity by analysing every word in the sentence (Check the parsed text for the ambiguity.
4. Establish the relationship between the neighboring words in the text using ontology information.
5. Resolve ambiguity with respect to sense closeness with context words (by applying suitable rules.)

The Rules are written using grammatical & relational information of Marathi language. The rules are automatically generated and applied.

Two classes of rules are framed.

1. Word Rules
2. Sentence Rules

Rules are applied to each word irrespective of its position and type. (position refers to the place of a word in the given text and type refers to the grammatical category of the word.)

Check for ambiguity module of phase II checks for the following result:

i) If Multiple possible sense list is obtained, it indicates presence of multiple ontology.

ii) For a single word, multiple rules may provide answers indicating ambiguity.

The GUI of our system is shown in figure 1 below. It includes buttons for all rules framed. It also shows the ambiguous words list. The buttons are categorized in word rule and sentence rule. The buttons indicating parsers of Noun, Verb, Adjective and Adverb are also available on GUI. It shows the time required for processing the sentence/ word in Milliseconds. The GUI provides keyboard for entering Marathi text for checking ambiguity. Figure 1 provides snapshot of GUI of our system.

Every word's POS and morphological details with ontological information is displayed on screen. The possible senses of each word are shown on GUI.

The words having multiple senses are displayed as senses

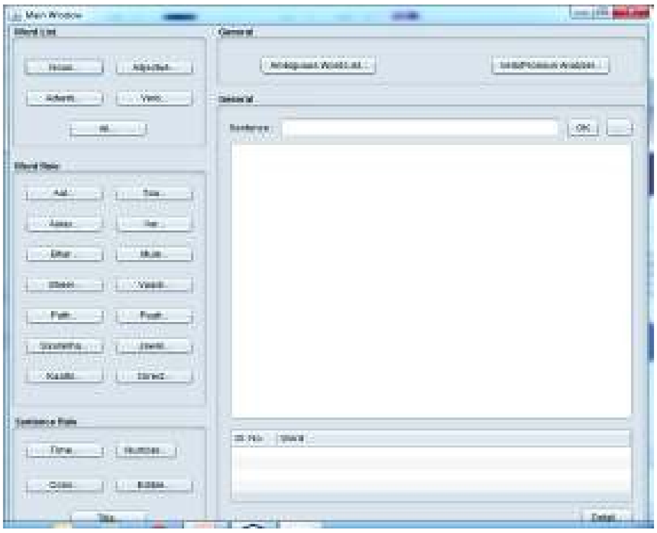


Figure 1: GUI design of our system

Word Rules: Some words are not directly available in dataset because they are attached with suffixes with some characters and therefore it is necessary to separate suffix & root word of these words. Words which are suffixed with special words have different sense as compared to the sense of their root word and therefore it is necessary to apply word sense rules to identify the correct sense of words.

Sentence Rules : In a sentence, there are more than one words. Here for each word of sentence we analyze that word belongs to which sense. Sometimes, a word's exact meaning is dependent on adjacent words. Hence the correct meaning of such words is not obtained using Word Sense Rules of Phase II. For example words like pandharaa ranga (white colour), aajaparyanta (untill today) etc.

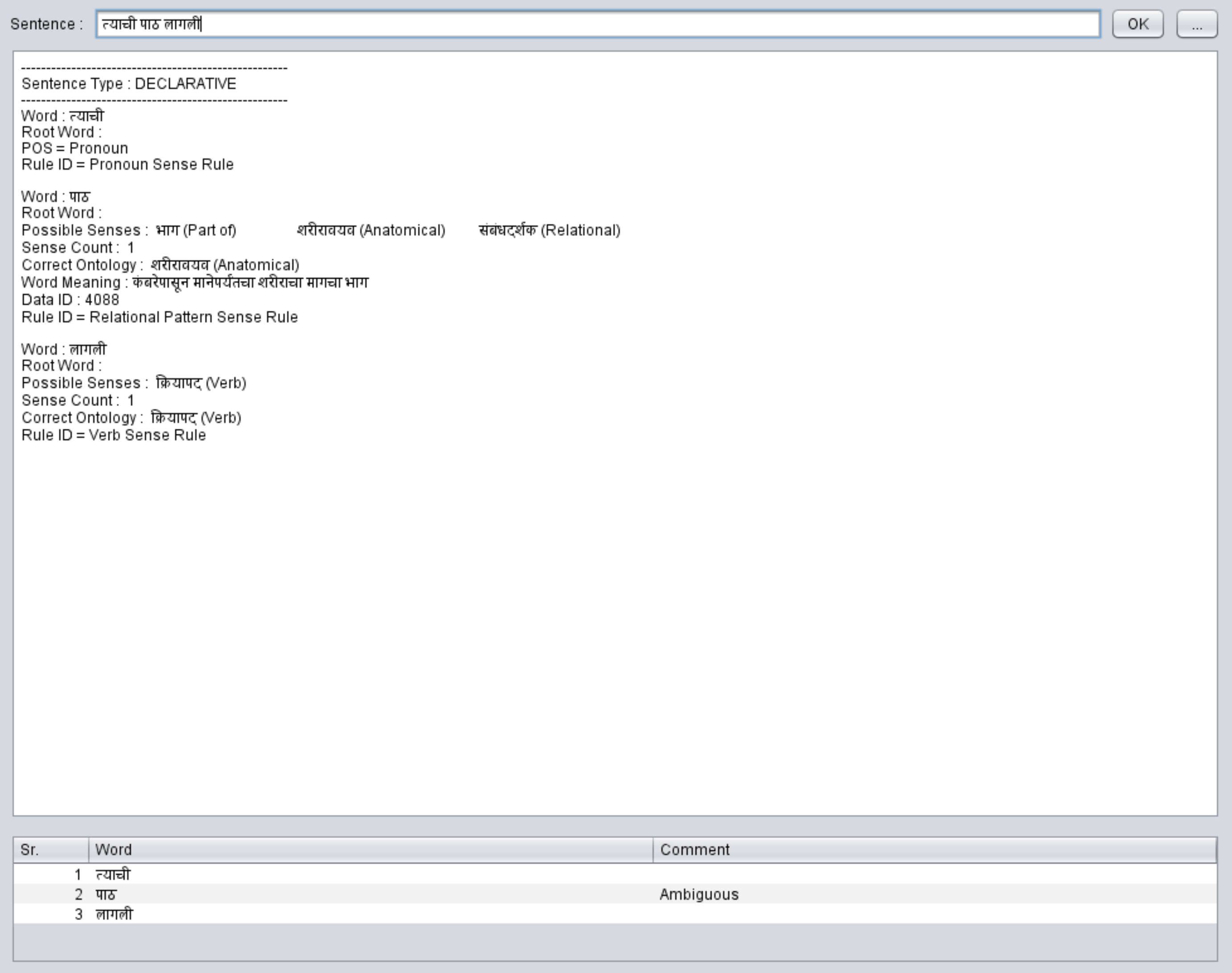
When more than one words of a sentence are used to set correct sense of the sentence, then the rule is called as Sentence Sense Rule.

4.2 Output of the system

The system developed is a knowledge based system which uses Wordnet 1.3 ontologies for Marathi language. The WSD system which is capable of performing Disambiguation for the Marathi language at word level, i.e. it resolves lexical and semantic ambiguity.

The output of the system i.e. disambiguation process is represented in following snapshots, the system GUI consist of number of rule list like word rule, sentence rule, pattern rule; word list as per grammatical category.

In the output word details are displayed like:
Root word
Pos tag of word
Possible senses
Correct senses
ontology reference etc.



Sentence : त्याची पाठ लागली

```
-----------------------------------------------------
Sentence Type : DECLARATIVE
-----------------------------------------------------
Word : त्याची
Root Word :
POS = Pronoun
Rule ID = Pronoun Sense Rule

Word : पाठ
Root Word :
Possible Senses :  भाग (Part of)          शरीरावयव (Anatomical)       संबंधदर्शक (Relational)
Sense Count :  1
Correct Ontology :  शरीरावयव (Anatomical)
Word Meaning : कंबरेपासून मानेपर्यंतचा शरीराचा मागचा भाग
Data ID : 4088
Rule ID = Relational Pattern Sense Rule

Word : लागली
Root Word :
Possible Senses :  क्रियापद (Verb)
Sense Count :  1
Correct Ontology :  क्रियापद (Verb)
Rule ID = Verb Sense Rule
```

Sr.	Word	Comment
1	त्याची	
2	पाठ	Ambiguous
3	लागली	

Figure 2: Snapshot of output

5. CONCLUSION

Thus the system discussed in this paper is a knowledge based solution to the problem of word sense disambiguation in Marathi language. At present WSD system is working at word level. The system accuracy is around 92% which include disambiguation of nouns, adjectives and verbs in the given Marathi language text. The limitation of the system is that it can only identify and resolve word level ambiguity.

ACKNOWLEDGEMENT

We wish to express gratitude towards IIT Mumbai CFILT experts for the valuable guidance extended in the work of WSD carried out by us in Marathi.

REFERENCES

[1] www.oxforddictionary.com

[2] Dixit, Veena, Dethe, Satish and Joshi, Rushikesh K. (2006), Design a nd Implementa tion of a Morphology-ba sed Spellchecker for Ma ra thi, a n Indian Language, In Special issue on Human, Language Technologies as a challenge for Computer Science and Linguistics. Part I. 15, pages 309–316. Archives of Control Sciences.

[3] http://www.glottopedia.org/index.php/Ambiguity,_Polys emy_and_Vagueness

[4] John Lyons, "Introduction to theoretical linguistics, Press Syndicate of University of Cambridge-digital printing 2001(Google Books:)

[5] Diane D. Bornstein, "An Introduction to Transformational Grammar" Book published by, University press of America) (Google Books)

[6] Lesk, M, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone", In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, pages 24-26, New York, NY, USA. ACM, (1986)

[7] Walker D. and Amsler R. 1986, "The Use of Machine Readable Dictionaries in Sublanguage Analysis", in Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press, pp. 69-83, 1986.

[8] Agirre, Eneko & German Rigau. 1996, "Word sense disambiguation using conceptual density", in Proceedings of the 16th International Conference on

Computational Linguistics (COLING), Copenhagen, Denmark, 1996

[9] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[10] http://www.cfilt.iitb.ac.in/wordnet/webmwn/

[11] Rohit Sharma, “Word Sense Disambiguation for Hindi Language”, Thesis dissertation, Thapar University, Patiala. July 2008.

[12] Saif Mohammad, “Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation”, Thesis dissertation, University of Minnesota, July 2003.

[13] Yarowsky, David, "Unsupervised word sense disambiguation rivaling supervised methods", Proceeding of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), Cambridge, MA, 189-196, 1995.

[14] Philip Resnik, "Selectional preference and sense disambiguation", Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics:Why, What, and How, 1997, pages 52–57.

[15] Jun Fu Cai, Wee Sun Lee, "Improving Word Sense Disambiguation Using Topic Features", Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1015–1023, Prague, June 2007. http://anthology.aclweb.org/D/D07/D07-1.pdf#page =1049

[16] D. Lin, "Dependency-based evaluation of minipar", Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation, Granada, Spain, May, 1998.

[17] Jacobs, Paul S, "Trump: A transportable language understanding program", International Journal of Intelligent Systems, 7:245-276, March 1992